Contents lists available at ScienceDirect

Engineering Applications of Artificial Intelligence

journal homepage: www.elsevier.com/locate/engappai

Forecasting crude oil price with a new hybrid approach and multi-source data

Yifan Yang, Ju'e Guo, Shaolong Sun [*], Yixin Li

School of Management, Xi'an Jiaotong University, Xi'an, 710049, China, China

ARTICLE INFO

Keywords:
Crude oil price forecasting
GSVI data
Kernel extreme learning machine
Herd behavior
Divide and conquer

ABSTRACT

Faced with the growing research toward crude oil price fluctuations influential factors following the accelerated development of Internet technology, accessible data such as Google search volume index (GSVI) are increasingly quantified and incorporated into forecasting approaches. In this study, we apply multi-scale data that including both traditional economic data and GSVI data reflecting macro and micro mechanisms affecting crude oil price respectively, so as to reduce the forecasting deviation and improve the forecasting accuracy at source. In addition, a new hybrid approach: K-means+KPCA+KELM based on "divide and conquer" strategy is proposed for deeply exploring the information of above multi-data so that improve monthly crude oil price forecasting accuracy. Empirical results can be analyzed from data and method levels. At the data level, GSVI data perform better than economic data in level forecasting accuracy but with opposite performance in directional forecasting accuracy because of "Herd Behavior", while hybrid data combined their advantages and obtain best forecasting performance in both level and directional accuracy. At the method level, the approaches with "divide and conquer" strategy gain a better forecasting performance, which demonstrates that "divide and conquer" strategy can effectively improve the forecasting performance.

1. Introduction

Crude oil, as the blood of the industry, plays an important role in the global economic market, whose price fluctuation has a significant impact on political and economic activities around the world (Ahmed et al., 2012). Policymakers who can accurately forecast crude oil price fluctuations can make prospective economic and political policies to gain advantages in a complex international environment. Therefore, based on the conviction that crude oil price forecasting can foster international prestige, crude oil price forecasting has become a topic that is never out of date, which has also received increased attention, and thus many practical measures and academic research are directed toward improving the accuracy of crude oil price forecasting. Influential factors of crude oil price except supply and demand such as economic growth, stock indexes and exchange rates also play an important role in crude oil price fluctuation (Baumeister et al., 2015; Ou et al., 2012; Singleton, 2014). It is therefore not surprisingly that scholars demonstrate that incorporating these influential factors into crude oil price forecasting frameworks can indeed improve forecasting accuracy.

With the accelerated development of Internet technology and a booming crude oil market, investors can gather more comprehensive information in real time through the Internet and make corresponding investment decisions in crude oil market. Meanwhile, the market of crude oil is one of the biggest commodity markets, short-term fluctuation of crude oil price could be caused by investors' investment decisions. Therefore, investor behavior has gradually become an emerging crude oil price influential factor that cannot be ignored. Behavioral Finance explains the contribution of investor behavior to crude oil price fluctuation psychologically and holds that asset price is not only determined by its intrinsic value but also influenced by the investors' behavior (Fama, 1998). Since human attention is a kind of scarce and limited resource, investors can only pay attention to the assets they concerned about. Thus, the psychological characteristics such as investor confidence can be revealed by the change of investor attention toward special asset, which can further influence the asset price fluctuation.

Although the measurement of investor attention is challengeable, search engines provide feasible solution to deal with this situation. Investors use search engines to search terms about assets they concerned, in such way the search volume and generated data are counted and collected, which can generate more objective and available measurements than those of traditional methods (Da et al., 2011). Search volume data, especially Google search volume index (GSVI) contributes to analyzing and forecasting various social and economic behaviors, such as disease surveillance, macro-economy index forecasting and tourism management. However, using GSVI to forecast crude oil price is still in its infancy, and there are few studies on how to improve the forecasting accuracy by combining GSVI with traditional macro-economic indicators.

In this article, multi-scale data are collected to reflect both macro and micro mechanisms that affect crude oil price, so as to reduce the

[*] Corresponding author.
E-mail address: sunshaolong@xjtu.edu.cn (S.L. Sun).

https://doi.org/10.1016/j.engappai.2021.104217
Received 30 March 2020; Received in revised form 15 November 2020; Accepted 27 February 2021
Available online 9 March 2021

forecasting deviation and improve the forecasting accuracy at source. Relying on the "related searches" function of Google search engine, forty GSVI terms are collected to represent the macroscopic influence mechanisms of crude oil price fluctuation. In addition, based on the existing literature, thirty-two traditional economic factors are selected to represent the microscopic influence mechanisms of crude oil price fluctuation.

Massive exogenous variables may lead to the "curse of dimensionality", which threaten the forecasting accuracy. However, K-means+ KPCA+KELM as a new hybrid approach can turn this situation around in the following reasons: firstly, K-mean method can divide influential variables into certain clusters based on its "divide and conquer" strategy. Secondly, following the principle of nonlinear dimensionality reduction, kernel principal component analysis (KPCA) combines the variables to composite indexes, which solves the overfitting problem while retaining most of the information. Finally, after dimension reduction of variables in each cluster, the composite indexes of all the clusters are mixed and kernel extreme learning machine (KELM) is adopted for final forecasting.

The remainder of this paper is organized as follows. The next section summarizes the related research, Section 3 introduces the methodologies and the framework of our proposed new hybrid approach. The empirical results are outlined in Section 4. After a discussion of the results, Section 5 concludes this study.

2. Literature review

2.1. Forecasting with search volume data

Prior research finds that GSVI data contributes to analyzing and forecasting various social and economic behaviors. In the field of disease surveillance, Ginsberg et al. (2009) used GSVI data to build an influenza epidemic forecasting model, which can forecast the intensity and timing of flu outbreaks one to two weeks in advance. Araz et al. (2014) used GSVI data to forecast Influenza-Like-Illness (ILI)-related emergency department visits in Omaha. Song et al. (2014) found that there is a significant positive correlation between GSVI of stress and the number of suicides in Korea. In the area of macro-economy, Smith (2016) highlighted there is a strong correlation between GSVI related to unemployment and the unemployment rate, and he further added GSVI data in a MIDAS regression framework to forecast unemployment in the UK. Li et al. (2015b) demonstrated that GSVI data and Consumer Price Index (CPI) officially released by the Statistic Bureau of China have a strong correlation, and the MIDAS forecasting model with GSVI outperforms the benchmarks in the reduction of root mean square error (RMSE) over 30%. Goetz and Knetsch (2019) included GSVI data in bridge equation models for German GDP forecasting. Additionally, in the research of tourism management, Bangwayo-Skeete and Skeete (2015) investigated GSVI related to hotel and flights in forecasting tourism demand of Caribbean by Autoregressive Mixed-Data Sampling (AR-MIDAS) models. Sun et al. (2019) forecast tourism arrivals of Beijing by GSVI and Baidu search volume index (BSVI) data, which outperforms the benchmarks without search engine data. Clark et al. (2019) applied GSVI data to forecast the tourism arrivals of U.S. National Parks and get a similar result.

However, using GSVI to forecast crude oil price is still in its infancy. Li et al. (2015a) used GSVI data to measure investor attention, so as to investigate the relationship among investor attention, trader position and weekly crude oil price. Guo and Ji (2013) applied GSVI data to analyze the impact of short- and long-run market concerns for crude oil price. Wang et al. (2018a) employed GSVI data to represent internet concern and propose a new framework to analysis internet concern for crude oil price volatility. These research only applied GSVI data of a few terms on average to build forecasting models, which cannot fully reflect investor attention and may lead to the bias of forecasting results. Han et al. (2017) first selected a wider set of GSVI of the terms to forecast weekly crude oil price, but they also linearly combined GSVI data into a composite index as the independent variable, which might cause omission of certain features. In a word, how to select comprehensive keywords and combine these keywords as efficient composite indexes are the core issues of using GSVI data for crude oil price forecasting.

2.2. Data preprocessing techniques of forecasting

As García et al. (2016) highlight: "Data preprocessing is a major and essential stage whose main goal is to obtain final data sets that can be considered correct and useful for further forecasting". Data preprocessing includes data cleaning, data integration, data transformation, and data reduction. We mainly focus on data reduction, and it will be introduced in detail later.[1]

There exist a wide range of techniques in the forecasting literature that aim at data reduction, among others: feature selection and feature extraction. Feature selection can recognize and eliminate as much irrelevant and redundant features as possible (Yan et al., 2006). Most studies of crude oil price forecasting tend to use feature selection for data reduction, because feature selection will retain its original features for better model interpretation. For instance, Miao et al. (2017) provided least absolute shrinkage and selection operator (LASSO) and stepwise regression model to determine the significant variables from 26 potential determinants of crude oil price. Lu et al. (2020) adopted dynamic Bayesian structural time series model to select 7 core factors from 415 independent variables for crude oil price analysis and forecasting. Some scholars also used similar feature selection method to refine exogenous variables (Fazelabdolabadi, 2019; Zhang et al., 2019a,b). However, feature selection only preserves valid variables by setting a certain threshold, so that a large amount of useful information is discarded, while feature extraction transforms the original feature space to a simpler one preserving more information (Yan et al., 2006). Feature extraction has already been applied for multivariate forecasting problems, such as wind speed forecasting (Sun et al., 2017), air quality forecasting (Sun and Sun, 2017) and stock market return forecasting (Zhong and Enke, 2017). but few studies adopted it for crude oil price forecasting with massive exogenous variables (Shin et al., 2013).

2.3. Forecasting approach of crude oil price

Previous crude oi price forecasting research is dominated by traditional econometric models and machine learning models. Traditional econometric models toward crude oi price forecasting include Autoregressive Integrated Moving Average (ARIMA) models (Mohammadi and Su, 2010; Moshiri and Foroutan, 2006), Generalized Autoregressive Conditional Heteroscedasticity (GARCH) family models (Hou and Suardi, 2012; Wei et al., 2010). Machine learning models vis-à-vis crude oil price forecasting include Neural network (NN) models (Huang and Wang, 2018; Movagharnejad et al., 2011; Yu et al., 2015). Support Vector Machine (SVM) models (Bisoi et al., 2019; Safari and Davallou, 2018; Yu et al., 2017), Wavelet-based models (Bisoi et al., 2019; Chai et al., 2018; Huang and Wang, 2018; Wang et al., 2018b). Both models have their disadvantages. Traditional econometric models have to make assumptions in advance and show poor performance in capturing nonlinear features, while machine learning models suffer from overfitting and parameters sensitive problems (Yu et al., 2008a,b).

Scholars then begin to apply deep learning model to improve crude oil price forecasting accuracy (Ghoddusi et al., 2019). Cen and Wang

[1] Our economic data are collected from commercial databases and government websites, and GSVI data are provided by Google Search, these data are clean, complete, and nearly normalized, it is thus data cleaning, data integration, and data transformation are not completely included in this study. Besides that, since the collected original data suffer from the curse of dimensionality because of massive exogenous variables, therefore we focus on the data reduction stage primarily.

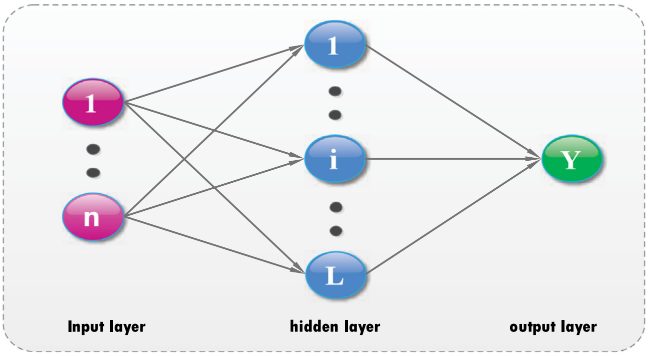


Fig. 1. The topological structure of ELM.

(2019) improved long short term memory (LSTM) model based on prior knowledge data transfer algorithm, and further applied it for WTI and Brent oil price forecasting. Li et al. (2019) used convolutional neural network (CNN) to explore the news headlines related to oil prices to improve crude oil price forecasting performance. Empirical researches demonstrated that the forecasting accuracy of deep learning models is significantly higher than that of traditional econometric models and other machine learning models. However, monthly crude oil price data cannot meet the big demand of data to conduct deep learning models.

Hybrid model based on its "divide and conquer" strategy can further improve forecasting accuracy with high-efficiency and thus in its prevalence (Hajirahimi and Khashei, 2019; Wang et al., 2005). To be specific, a series of decomposition algorithm is first used to decompose the raw data into different modes, then different models are used to forecast different modes according to their data characteristics, and finally the forecasting results are integrated. For example, Yu et al. (2008b) first applied an "EMD (Empirical Mode Decomposition)-FNN (Feedforward Neural Network)-ALNN (Adaptive Linear Neural Network)" hybrid approach to forecast daily crude oil price. Panigrahi and Behera (2017) proposed a hybrid ETS–ANN model to forecast sixteen time series respectively, where ETS stands error, trend and seasonal part of oral series. Tang et al. (2018) also utilized Ensemble Empirical Mode Decomposition (EEMD) and Random Vector Functional Link (RVFL) network for crude oil price forecasting.

However, such decomposition integration algorithms fit for single price series or forecasting contains a small amount of influential factors. Therefore, another framework is proposed to deal with the problem of crude oil price forecasting with a number of influential factors while retaining the advantages of "divide and conquer" strategy, which will be described in the next section.

3. Methodology

In this section, our proposed new hybrid approach K-means+KPCA+KELM and its adopted basic model are introduced. K-means and Kernel Principal Component Analysis (KPCA) are basic statistical learning methods, whose principle would not be presented in this study, more detail please refer to Kanungo et al. (2002) and Scholkopf et al. (1998).

3.1. Kernel extreme learning machine

Extreme learning machine (ELM), one type of effective single-hidden layer feedforward neural network (SLFN), has been widely used in many fields (Liu et al., 2020; Sun et al., 2020; Yu et al., 2016). ELM can generate randomly the weight and bias of a hidden layer, which need not be tuned anymore. After determining the number of hidden nodes and activation function, output weights can be obtained by matrix computations rather than iteration (Huang et al., 2006).

Given N samples $(x_i, y_i), x_i \in \Re^N, y_i \in \Re^N, i = 1, 2, \dots, N$, for a typical ELM with n input neurons, L hidden neurons and one output neurons (shown in Fig. 1), one can define the output matrix of ELM as:

$$Y = \begin{bmatrix} y_{1j} \\ y_{2j} \\ \vdots \\ y_{mj} \end{bmatrix}_{m\times N} = \begin{bmatrix} \sum_{i=1}^{L} \beta_{i1} h\left(\omega_i x_j + b_i\right) \\ \sum_{i=1}^{L} \beta_{i2} h\left(\omega_i x_j + b_i\right) \\ \vdots \\ \sum_{i=1}^{L} \beta_{im} h\left(\omega_i x_j + b_i\right) \end{bmatrix}_{m\times N}, (j = 1, 2, \dots, N) \quad (1)$$

where β is the output weights matrix between the hidden layer and the output layer, $h(\cdot)$ is the activation function of the hidden layer, $\omega_i = \left[\omega_{i1}, \omega_{i2}, \dots, \omega_{iN}\right]^T$ is the input weights matrix between ith input layers and hidden layers, and b is the biases matrix of the hidden layer.

We can also rewrite Eq. (1) as:

$$Y = H\beta, Y \in \Re^{N\times m}, \beta \in \Re^{N\times m} \quad (2)$$

where $H = H(\omega, b) = h(\omega x + b)$ is the output matrix of the hidden layer.

The value of input weights and biases are randomly assigned rather than being tuned. Thus, the output weights are the only unknown parameters, which can be calculated by the ordinary least square (OLS), the result can be written as:

$$\hat{\beta} = H^{\dagger} Y \quad (3)$$

where $H^{\dagger}$ is denoted as the Moore–Penrose generalized inverse of the output matrix.

Based on Ridge Regression Theory and Karush–Kuhn–Tucker (KKT) theorem, we can also add a positive penalty term $1/C$ to recalculate β as:

$$\hat{\beta} = H^T \left(I/C + HH^T\right)^{-1} Y \quad (4)$$

Therefore, the output function of ELM can be presented as:

$$\mathrm{f}(x) = H\hat{\beta} = HH^T \left(I/C + HH^T\right)^{-1} Y \quad (5)$$

The main idea of KELM is to replace the activation function of ELM as a kernel function according to Mercer's conditions, the output function of KELM can be presented as:

$$\mathrm{f}(x) = H\hat{\beta} = \begin{bmatrix} k\left(x, x_1\right) \\ k\left(x, x_2\right) \\ \vdots \\ k\left(x, x_n\right) \end{bmatrix}^T \left(I/C + HH^T\right)^{-1} Y \quad (6)$$

where $k(x, x_i)$ represents the kernel function.

3.2. The framework of our proposed new hybrid approach

In this paper, based on "divide and conquer" strategy, a new hybrid approach named K–means-KPCA–KELM is proposed to forecast crude oil price. It is formulated by following three steps. Fig. 2 illustrates the framework of our proposed new hybrid approach.

Step 1: Data fusion. Collecting the GSVI series of oil-related terms and filter out the irrelevant and unrelated terms, then merge the remaining GSVI series with other economic series as independent variables series.

Step 2: Dimension reduction. K-means method divides independent variables series into k clusters in terms of their correlation degree. For each cluster, KPCA is adopted to reduce data dimensions and obtain low dimension features, which reduces the data complexity while retaining as much useful information as possible.

Step 3: Forecasting. Combining the above features as the input matrix of KELM to forecast crude oil price, and adopting a series of evaluation criteria to evaluate the performance of our proposed approach.

Generally speaking, our K-means+KPCA+KELM approach contributes to optimizing data preprocessing. Due to massive exogenous variables, data reduction is unavoidable (García et al., 2016). We add a clustering operation before dimension reduction to retain as much

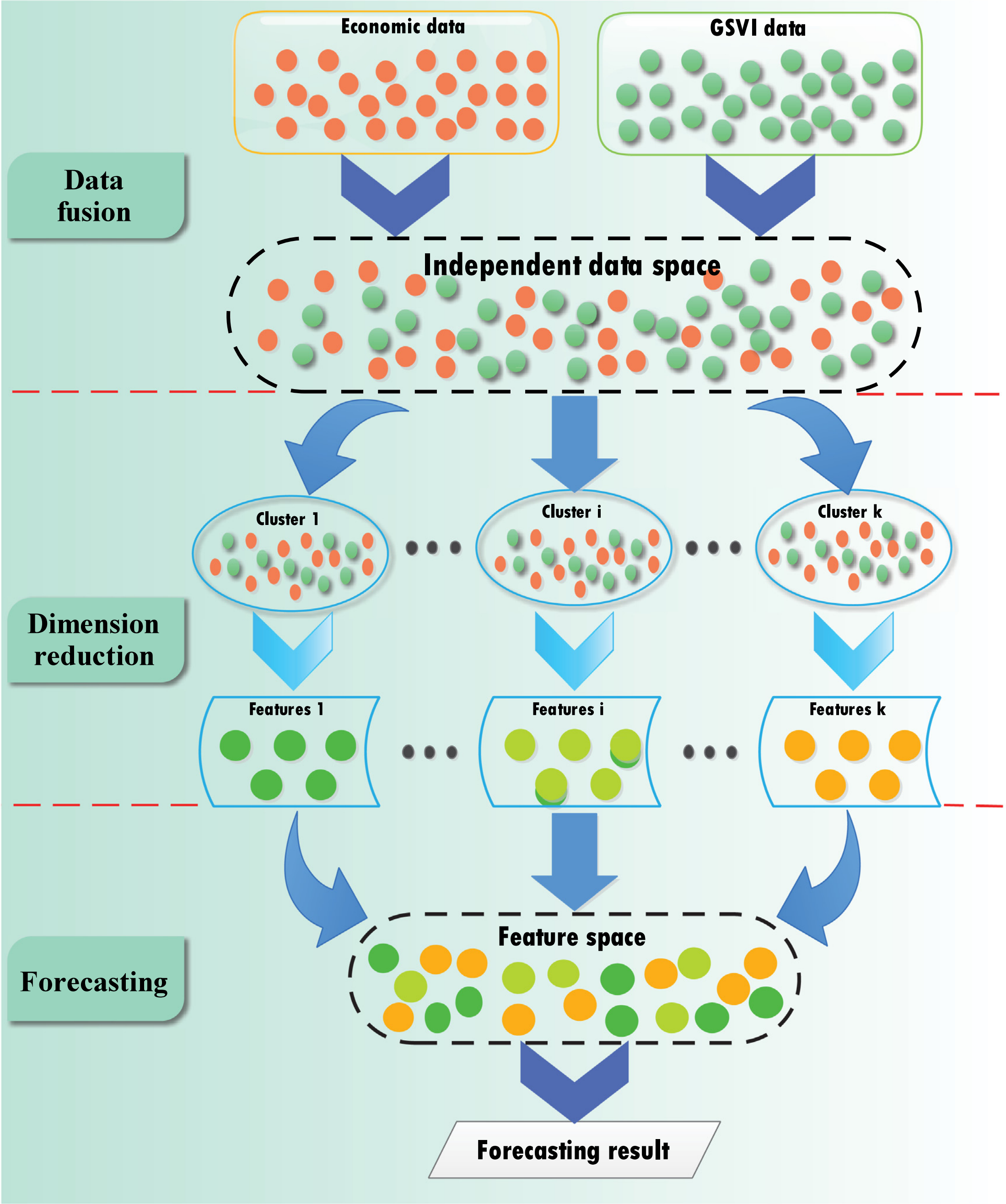


Fig. 2. The framework of our proposed new hybrid approach.

forecasting information as possible while realizing data reduction on the basis of "divide and conquer" strategy. Our approach has its priority in the two ways. First, the forecasting information might distribute in several principle components (PCs) under directly dimension reduction. The approach would group exogenous variables according to their similar characteristics, and thus forecasting information tends to concentrate in fewer heading PCs during dimension reduction in each group. By selecting heading PCs, the most of forecasting information in the original data can be retained and grasped for further forecasting. Second, the clustering operation avoid the curse of dimensionality to some extent since we only need to carry out a series of relatively low-dimensional dimension reduction operations, and merge corresponding PCs to obtain the final low-detention data. As a consequence, computational complexity can be significantly reduced.

4. Empirical study

4.1. Data collection

In this paper, the monthly West Texas Intermediate (WTI) crude oil spot price series(shown in Fig. 3) extracted from Wind Database (http://www.wind.com.cn/) is used as the dependent variable. In addition, economic dataset and GSVI dataset range from January 2004 to December 2018 are collected as the independent variables. Then we divide those datasets into two parts: the train datasets range from January 2004 to December 2017 and the test datasets range from January 2018 to December 2018. The following subsections describe the above datasets.

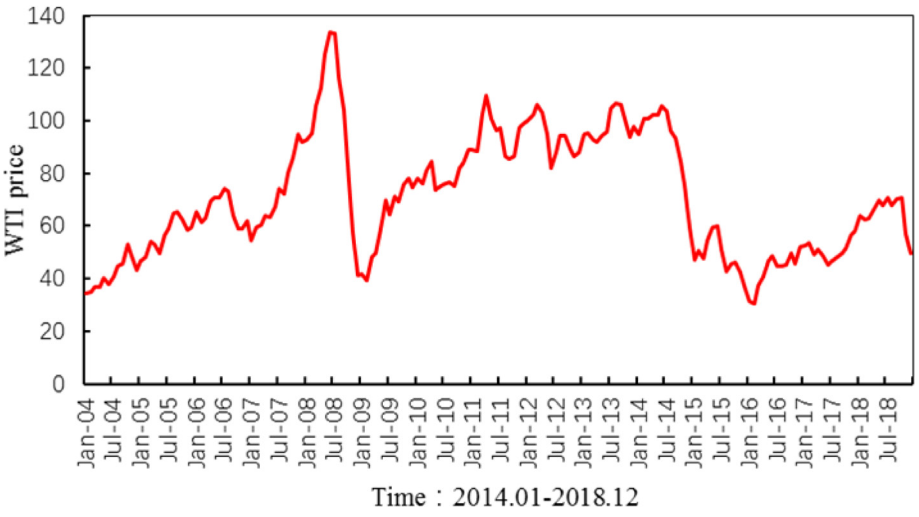


Fig. 3. Monthly WTI crude oil spot price.

4.1.1. *Economic dataset*

Supply, demand, and inventory are called three cornerstones that affect crude oil price. Thus, we first consider economic variables related to these basic influencing factors. In this paper, the supply-related variables include crude oil production capacity, refining capacity, consumption structure and replacement cost. The demand-related variables are comprised of the volume of energy consumption and some indexes about global economic development. The inventory-related variables are oil stocks. Moreover, since crude oil price interacts with other economic and financial market activities, relevant variables are added into independent variables as well, such as the monetary market index, commodity market index, and stock market index. Economic dataset is described in Table 1.

4.1.2. *GSVI dataset*

In this study, we use Google search volume index (GSVI), generated by a public tool (https://trends.google.com/) of Google Inc., as a proxy variable of investor attention for three reasons. First, Google search is the most popular search engine that can offer a huge amount of free and available online data. Second, GSVI consists of normalized structural data range from zero to 100, where zero refers that search volume is below a certain threshold, and 100 refers to a higher limit. Third, since this paper focuses on international crude oil price forecasting instead of Chinese domestic crude oil price forecasting, GSVI is more suitable than other search volume indexes such as Baidu search volume index due to its worldwide adoption.

Extracting useful information from a large amount of data is challengeable, we thus apply a three-stage process to refine useful search terms based on these paper (Afkhami et al., 2017; Han et al., 2017).

Firstly, we build an oil-related terms seed-set based on the following aspects: (1) add the terms directly related to crude oil price such as “oil price”, “oil demand” and “oil supply”; (2) add the terms related to other economic indexes and variables such as “gold price and “GDP”; (3) add oil-related terminologies from the glossary of the Colorado Oil and Gas Conservation Commission (COGCC) and some renewable energy terms; (4) add attention terms with a tendency to fear such as “crisis” and “bankrupt”. Secondly, we search the terms of our seed-set in Google Trend and iteratively set recommended terms as second-round search terms. This process is repeated until there are no new terms in the recommended list. Thirdly, we estimate the degrees of relevance with crude oil price series for the above terms by Granger causality test and filter out the terms whose p-value over 0.1. Finally, a set of 40 GSVI terms is built in alphabetical order (shown in Table 2).

4.2. *Performance evaluation criteria*

We apply mean absolute percentage error (MAPE), root mean square error (RMSE) and directional accuracy (DA) to evaluate the forecasting accuracy of our proposed new hybrid approach from the level and directional aspect respectively:

$$MAPE = \frac{1}{N}\sum_{t=1}^{N}\left|\frac{y(t)-\hat{y}(t)}{y(t)}\right| \times 100\% \tag{7}$$

$$RMSE = \sqrt{\frac{1}{N}\sum_{t=1}^{N}\left(y(t)-\hat{y}(t)\right)^2} \tag{8}$$

$$DA = \frac{1}{N}\sum_{t=1}^{N} d(t) \times 100\% \tag{9}$$

$$d(t) = \begin{cases} 0 & if \quad (y(t+1)-y(t))(\hat{y}(t+1)-y(t)) < 0 \\ 1 & if \quad (y(t+1)-y(t))(\hat{y}(t+1)-y(t)) \geq 0 \end{cases} \tag{10}$$

where N denotes the number of observations, $y(t)$ and $\hat{y}(t)$ denote the actual crude oil price and forecasting crude oil price respectively.

MAPE and RMSE measure the level accuracy, the smaller the MAPE/RMSE, the better the level performance. DA measures the directional accuracy, the higher the DA, the better the directional performance.

Moreover, we introduce the improvement rate (IR) to test the superior forecasting ability of our proposed new hybrid approach compared with its benchmarks:

$$IR_{MAPE} = -\frac{MAPE_A - MAPE_B}{MAPE_B} \times 100\% \tag{11}$$

$$IR_{RMSE} = -\frac{RMSE_A - RMSE_B}{RMSE_B} \times 100\% \tag{12}$$

$$IR_{DA} = \frac{DA_A - DA_B}{DA_B} \times 100\% \tag{13}$$

where approach A represents the proposed approach and approach B denotes the benchmark. When approach A outperforms approach B, the value of IR is positive and vice versa.

4.3. *Benchmarks and parameters setting*

A series of single models and hybrid models are selected as benchmarks to test the superior forecasting ability of K-means–KPCA–KELM

Table 1
Description of the economic dataset.

First-class index	Second-class index	Variables	Data Source
Supply	Production	Crude Oil Production, Total OPEC	EIA
		Crude Oil Production, Total Non-OPEC	EIA
		Crude Oil Production, World	EIA
	Consumption structure	Henry Hub Natural Gas Spot Price	EIA
		Rest US tight oil	EIA
	Technology	WTI-Brent spot price spread	EIA
		WTI crack spread: actual value	EIA
		Brent crack spread: actual value	EIA
Demand	Consumption	Petroleum Consumption, Total OECD	EIA
		China oil import	Wind database
	Global economic development	Fed fund effective	FRB
		Kilian Global economic index	https://sites.google.com/site/lkilian2019
		US: CPI index: seasonally adjusted	Wind database
		US: CPI: energy: seasonally adjusted	Wind database
		US: PPI: manufacturing sector total	Wind database
		US: PPI: mining sector total	Wind database
		EU 28 Countries: PPI	Wind database
		US PMI index	Wind database
Inventory	Inventory	Petroleum Stocks, Total OECD	EIA
		Crude Oil Stocks, Total	EIA
		Crude Oil Stocks, SPR	EIA
		Crude Oil Stocks, Non-SPR	EIA
Market activity	Monetary Market	Real dollar index: generalized	The federal reserve
		Exchange rate of euro against US dollar	The federal reserve
	Stock market	S&P 500 Index	Wind database
		Dow Jones Industrial Index	Wind database
		NYSE Index	Wind database
		AMEX Index	Wind database
		NASDAQ index	Wind database
	Commodity market	COMEX: Gold: Future closing price	Wind database
		LME: Copper: Future closing price	Wind database
		Crude oil non-commercial net long ratio	CFCT

Table 2
GSVI data terms.

Alternative energy	Crude oil	Energy security	Great depression	Recession depression
Bankrupt	Crude price	Expensive	Greenhouse gases	Recession
Bankruptcy	Crude prices	Financial situation	Horizontal drilling	State of the economy
Brent crude oil futures	Current interest	Fossil fuel	Interest rates	U.S. economy
Brent crude	Economic issues	Frugal	Kerosene	Unemployment
Carbon footprint	Economic situation	Gamble	Natural gas price	West Texas Intermediate
Carbon intensity	Economy problems	Gas subsidy	Offshore drilling	WTI oil
Clean energy	Energy conservation	Going green	Oil price	WTI price

model. Firstly, five single models are adopted for univariate forecasting, and effective models are selected for further study through the comparison of performance evaluation criteria, where ARIMA is the most basic and popular econometric model, ANN and LSSVR are most popular machine learning models, ELM and KELM are our proposed forecasting models. Scholars have demonstrated the good performance of all these models in crude oil price forecasting. Secondly, the hybrid approaches extended by selected models are applied for multivariable forecasting. Taking KELM as an example, the first type of hybrid approaches only applies dimension reduction for data preprocessing (KPCA+KELM), while the other types of hybrid approach conduct clustering operation before dimension reduction (K-means+KPCA+KELM). We apply all the multivariable approaches on three different types of independent variables datasets: GSVI dataset, economic dataset and hybrid dataset (and hybrid dataset includes both GSVI dataset and economic dataset).

It is worth mentioning that both K-means+PCA+KELM and PCA+KELM obtain poor forecasting performances in our datasets. It is suggested that PCA is more suitable for linear problems but crude oil price series is non-linear, uncertain and dynamic, and it is therefore not surprising that PCA is not applied as a dimensional reduce method in this study.

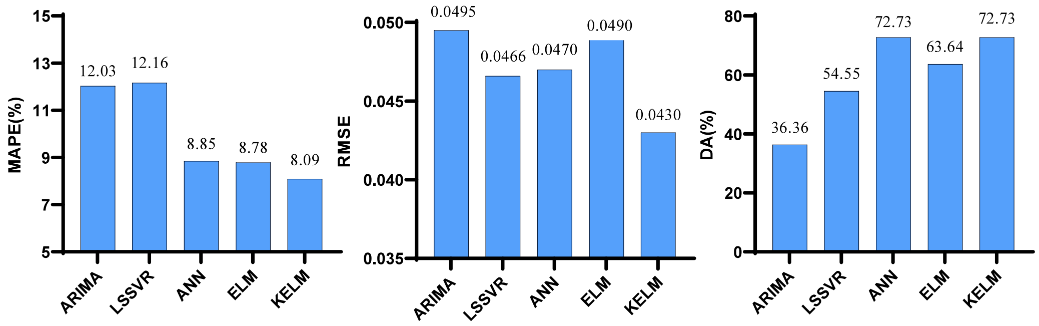
In this study, the optimum clustering numbers of K-means is determined as 3 according to "Elbow Criterion". The Gaussian kernel function is adopted in both KPCA and KELM as kernel function. The optimal lag of ARIMA is estimated by means of Akaike Information Criterion (AIC) and Schwarz Criterion (SC). The rest of the parameters are selected by trial and error testing by means of the minimization of Mean Absolute Error (MAE). All models are running by Matlab R2018a software on a server with 4 Core CPU of i5-4590 3.30 GHz, RAM size of 8 GB.

4.4. Empirical results

To test our forecasting approach, firstly, five single models are conducted to forecast monthly WTI crude oil spot price in order to find the best single forecasting model. Secondly, multivariable approaches (including our proposed new hybrid approach) are applied in economic dataset, GSVI dataset and hybrid dataset respectively. The results are interpreted from both data and method perspectives to demonstrate the superior forecasting ability of our proposed new hybrid approach with hybrid dataset. Departing from these, this paper also provides and explains several interesting phenomena at last.

4.4.1. Forecasting performance comparison of single models

The forecasting performances of single models with the WTI crude oil spot price dataset in Fig. 4 shows that: (1) KELM has the best forecasting performance, followed by ELM and ANN, LSSVR and ARIMA rank the last. (2) ARIMA, as a traditional econometric model, has poor performance in capturing the nonlinear and dynamic features of crude oil price series, thus the performance of ARIMA is worse than



(a) MAPE of single model (b) RMSE of single model (c) DA of single model

Fig. 4. Forecasting performance comparison of different single models.

those of the other machine learning models especially in directional forecasting accuracy. (3) As stated in the NFL (No free Lunch) theorem, no model can obtain good performance in all problems. SVM has strong performance in previous forecasting problems, but shows poor performance in our research context. (4) The performances of ELM and ANN are quite similar to KELM via both level evaluation criteria and directional evaluation criteria, which suggest that models based on neural network are more appropriate for our forecasting issue.

Therefore, KELM, ELM and ANN are considered as the best single models for univariate forecasting and they are selected as the basic models for our hybrid multivariable approaches in the following steps.

4.4.2. Forecasting performance comparison of multivariable approaches

The forecasting performances of multivariable approaches in three different types of datasets are discussed as follows, Fig. 5 shows the performance comparison results of six multivariable approaches in different datasets. It is shown that our proposed K-means+KPCA+KELM approach with hybrid dataset has the lowest MAPE: 5.44%, lowest RMSE: 0.0311 and highest DA: 90.91%. In general, the multivariable approaches are more efficient than single models. Because the independent variables of multivariable approaches contain a lot of information to capture more features of crude oil price. Moreover, as the results display, the performance of KELM is slightly better than ELM and ANN in all groups, it is therefore reasonable to select ANN, ELM and KELM as basic single models.

Next, we apply IR criteria to analyze the empirical results from the data and method perspective respectively, which further support the superior forecasting ability of hybrid dataset and our proposed new hybrid approach.

Table 3 shows the improvement rate of three evaluation criteria for different datasets, where E, G and H represent economic dataset, GSVI dataset and hybrid dataset respectively. It is clear that: (1) For each group, the IR values of $H \rightarrow E$ and $H \rightarrow G$ are positive in both level and directional performance evaluation criteria, which reveals that hybrid dataset contributes significantly more than only economic dataset or GSVI dataset in crude oil price forecasting. (2) For each group, the IR values of $E \rightarrow G$ are positive in level performance evaluation criteria while the values are negative in directional performance evaluation criteria, which shows that economic dataset contributes significantly more than GSVI dataset in level forecasting but with opposite performance in directional forecasting.

Table 4 displays the contribution of clustering operation in dimension reduction, in which Approach 1 Approach 2 and Approach 3 represent KPCA+ANN, KPCA+ELM and KPCA+KELM, and Approach 4, Approach 5 and Approach 6 refer to approaches that combined K-means method with the above models respectively. For each group, Approach 4~Approach 6 outperform Approach 1~Approach 3 respectively according to the positive IR values. It is obvious that the approaches with K-means method not only obtain the highest level forecasting accuracy (via the positive IR_{MAPE} and IR_{RMAE} criteria) but also acquire the best directional forecasting performance (via the positive IR_{DA} criteria), which indicates that clustering operation contributes a lot for forecasting performance improvements.

Table 3
IR between different datasets.

Approaches	Datasets	IR_{MAPE}(%)	IR_{RMSE} (%)	IR_{DA}(%)
KPCA+ANN	$E \rightarrow G$	13.66	4.64	0.00
	$H \rightarrow G$	41.93	30.95	33.33
	$H \rightarrow E$	32.75	27.58	33.33
K-means+KPCA+ANN	$E \rightarrow G$	22.51	20.69	0.00
	$H \rightarrow G$	25.02	30.02	16.67
	$H \rightarrow E$	3.32	11.77	16.67
KPCA+ELM	$E \rightarrow G$	21.20	13.19	−14.29
	$H \rightarrow G$	44.04	29.21	14.29
	$H \rightarrow E$	28.99	18.46	33.33
K-means+KPCA+ELM	$E \rightarrow G$	22.89	20.19	−12.5
	$H \rightarrow G$	35.85	21.87	0.00
	$H \rightarrow E$	16.81	2.11	14.29
KPCA+KELM	$E \rightarrow G$	30.65	24.50	−22.22
	$H \rightarrow G$	47.70	45.01	0.00
	$H \rightarrow E$	24.59	27.17	28.57
K-means+KPCA+KELM	$E \rightarrow G$	24.92	34.74	−11.11
	$H \rightarrow G$	40.32	37.21	11.11
	$H \rightarrow E$	20.51	3.79	25.00

Note: E represents economic dataset; G represents GSVI dataset; H represents hybrid dataset.

Table 4
IR between different approaches.

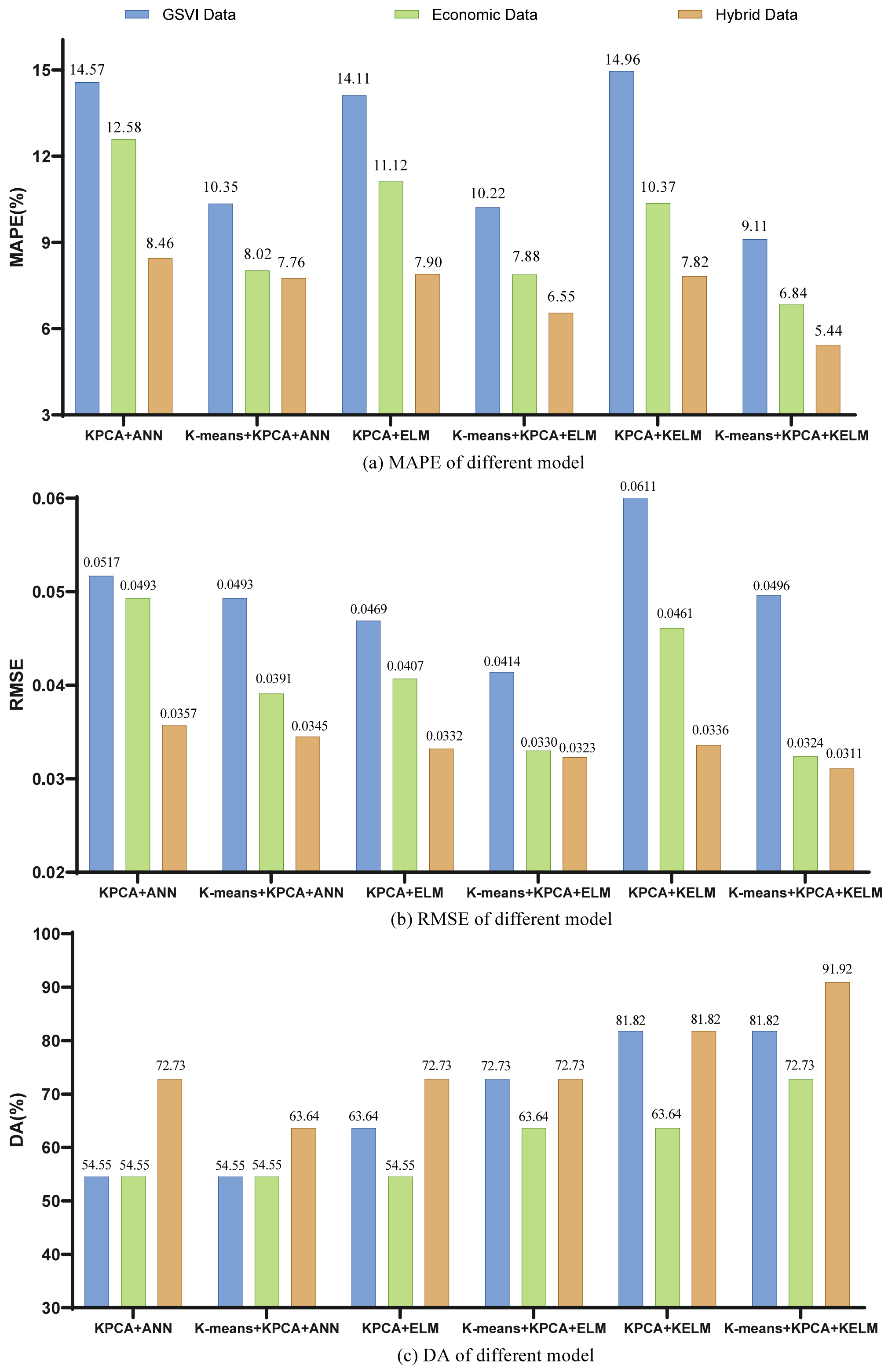
Datasets	Approaches	IR_{MAPE}(%)	IR_{RMSE} (%)	IR_{DA}(%)
GSVI Dataset	$Approach4 \rightarrow Approach1$	28.96	4.64	0.00
	$Approach5 \rightarrow Approach2$	27.61	11.79	14.29
	$Approach6 \rightarrow Approach3$	39.10	18.85	0.00
Economic Dataset	$Approach4 \rightarrow Approach1$	36.25	20.69	0.00
	$Approach5 \rightarrow Approach2$	29.16	18.90	16.67
	$Approach6 \rightarrow Approach3$	34.07	29.86	14.29
Hybrid Dataset	$Approach4 \rightarrow Approach1$	8.27	3.36	−12.50
	$Approach5 \rightarrow Approach2$	17.02	2.64	0.00
	$Approach6 \rightarrow Approach3$	30.51	7.35	11.11

Note: Approach 1: KPCA+ANN; Approach 2: KPCA+ELM; Approach 3: KPCA+KELM; Approach 4: K-means+KPCA+ANN; Approach 5: K-means + KPCA + ELM; Approach 6: K-means+KPCA+KELM.

4.4.3. Discussions

According to the above performance comparisons of both single models and multivariable approaches, it is clear that K-means+KPCA+KELM outperforms other benchmarks in both level and directional accuracy, and hybrid dataset obtains better forecasting performance than single economic dataset or GSVI dataset via both level and directional evaluation criteria. Moreover, we would like to make brief explanations of two interesting phenomena found in Tables 3 and 4:

(1) As Table 3 shown, hybrid dataset has the best performance in level accuracy, followed by economic data and GSVI dataset ranks



(a) MAPE of different model

(b) RMSE of different model

(c) DA of different model

Fig. 5. Performance comparison of different hybrid approaches.

the last, while hybrid dataset has the best performance in directional accuracy, followed by GSVI dataset and economic dataset ranks the last. Economic dataset includes macroeconomic influencing factors and determines the real value of crude oil, which significantly reflect the trend and period components of crude oil price. Besides, economic data is not sensitive to sudden events such as wars and abnormal climate that lead to the short-term fluctuations of crude oil price. GSVI dataset is composed of the proxy variables of investor attention so it can reveal the investor behavior. In the capital market, a single investor always acts according to the actions of other similar investors, buys when others buy, and sells when others sell, this phenomenon is called “Herd Behavior”. Thus, investors are sensitive to sudden events and react quickly, thus GSVI dataset has a strong ability to capture the short-term fluctuation components of crude oil price, which contributes a lot for directional forecasting ability. It is nevertheless true that investors tend to exaggerate the degree of crude oil price fluctuations because of ''Herd Behavior'', which reduces the level forecasting accuracy. However, hybrid dataset, combined economic dataset with GSVI dataset, can not only capture the trend, period components, but also capture the short-term fluctuations components of crude oil price without exaggeration. In brief, economic dataset tends to improve level forecasting accuracy while GSVI dataset tends to improve directional forecasting accuracy, and hybrid dataset combines their advantages to get the best performance in both level and directional forecasting accuracy.

(2) As Table 4 shown, the approaches with K-means method perform better than corresponding approaches without K-means. Based on ''divide and conquer'' strategy, our proposed new hybrid approach first divides the input data into k clusters, then individually reduce dimensions for each cluster, and thirdly group these low dimension features as new input data for the forecasting model. Compared with the direct dimension reduction method, ''divide and conquer'' strategy is more refined and effective, which can discover the unique properties for different components of origin series.

5. Conclusions

In this paper, we combined economic dataset with GSVI dataset as independent variables for crude oil price forecasting, where the two datasets reflect the impact of macro-economic variables and micro-individual behavior respectively. In order to fully exploit and utilize the information of above dataset, we proposed a new hybrid approach combined with K-means, KPCA and KELM, where K-means method is applied to divide independent variables into k clusters according to their correlation degree, KPCA is adopted to map independent variables into low dimensional space, KELM is employed for final crude oil price forecasting. Our empirical results show that our proposed new hybrid model significantly outperforms other benchmarks in both level and directional accuracy for each dataset and hybrid dataset performs better than other datasets in both level and directional accuracy for every approach.

Based on these results, the contribution of our work is threefold. Firstly, compared with the traditional econometric model, our proposed single model KELM has a strong ability in capturing the nonlinear and dynamic features of crude oil price and outperform other single models. Secondly, GSVI dataset has a strong ability to capture the short-term fluctuations components of crude oil price, due to the existence of the ''Herd Behavior'', GSVI dataset often exaggerate the degree of those fluctuations, while economic dataset reflects more trend and period components of crude oil price and less short-term fluctuations components. Our proposed hybrid dataset, composed by economic dataset and GSVI dataset, combines their advantages to capture trend and period as well as short-terms fluctuations components of crude oil price. Thirdly, based on ''divide and conquer'' strategy, this paper performs a clustering operation before reduce dimension, which prefer to discover more information about crude oil price fluctuations and improve forecasting accuracy.

It is suggested that our proposed new hybrid approach based on ''divide and conquer'' strategy and multi-scale data fusion especially GSVI data can be applied as independent variables to obtain a better forecasting performance in other complex forecasting issues such as power load or consumption forecasting, traffic flow forecasting and PM2.5 concentration forecasting.

However, since this paper applies the most common Gaussian functions as the kernel function in KPCA and KELM, it is suggested that other alternatives functions substitute for Gaussian can further improve the forecasting accuracy. In addition, some parameters in this paper are determined by trial and error testing, which is time-consuming and not suitable for large-scale data processing. Hence, a more appropriate and time-saving method to select optimal parameters should be exercised in future research.

CRediT authorship contribution statement

Yifan Yang: Conceptualization, Software, Formal analysis, Writing - original draft. **Ju'e Guo:** Conceptualization, Resources, Funding acquisition. **Shaolong Sun:** Methodology, Formal analysis, Funding acquisition. **Yixin Li:** Methodology, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This research work was partly supported by the National Natural Science Foundation of China under Grants No. 71774130 and No. 71988101 and the Fundamental Research Funds for the Central Universities, China under Grant No. xpt012020022. Authors would like to express their sincere appreciation to the editor and the two independent referees in making valuable comments and suggestions to this paper. Their comments and suggestions have improved the quality of the paper immensely.

References

Afkhami, M., Cormack, L., Ghoddusi, H., 2017. Google search keywords that best predict energy price volatility. Energy Econ. 67, 17–27.

Ahmed, H.J.A., Bashar, O.H.M.N., Wadud, I.K.M.M., 2012. The transitory and permanent volatility of oil prices: What implications are there for the US industrial production? Appl. Energy 92, 447–455.

Araz, O.M., Bentley, D., Muelleman, R.L., 2014. Using Google Flu Trends data in forecasting influenza-like-illness related ED visits in omaha. Nebraska. Am. J. Emerg. Med. 32, 1016–1023.

Bangwayo-Skeete, P.F., Skeete, R.W., 2015. Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. Tourism Manag. 46, 454–464.

Baumeister, C., Guerin, P., Kilian, L., 2015. Do high-frequency financial data help forecast oil prices? The MIDAS touch at work. Int. J. Forecast. 31, 238–252.

Bisoi, R., Dash, P.K., Mishra, S.P., 2019. Modes decomposition method in fusion with robust random vector functional link network for crude oil price forecasting. Appl. Soft Comput. 80, 475–493.

Cen, Z., Wang, J., 2019. Crude oil price prediction model with long short term memory deep learning based on prior knowledge data transfer. Energy 169, 160–171.

Chai, J., Xing, L.-M., Zhou, X.-Y., Zhang, Z.G., Li, J.-X., 2018. Forecasting the WTI crude oil price by a hybrid-refined method. Energy Econ. 71, 114–127.

Clark, M., Wilkins, E.J., Dagan, D.T., Powell, R., Sharp, R.L., Hillis, V., 2019. Bringing forecasting into the future: Using Google to predict visitation in US national parks. J. Environ. Manag. 243, 88–94.

Da, Z., Engelberg, J., Gao, P., 2011. In search of attention. J. Financ. 66, 1461–1499.

Fama, E.F., 1998. Market efficiency, long-term returns, and behavioral finance. J. Financ. Econ. 49, 283–306.

Fazelabdolabadi, B., 2019. A hybrid Bayesian-network proposition for forecasting the crude oil price. Financ. Innov. 5, http://dx.doi.org/10.1186/s40854-019-0144-2.

García, S., Luengo, J., Herrera, F., 2016. Tutorial on practical tips of the most influential data preprocessing algorithms in data mining. Knowl.-Based Syst. 98, 1–29.

Ghoddusi, H., Creamer, G.G., Rafizadeh, N., 2019. Machine learning in energy economics and finance: A review. Energy Econ. 81, 709–727.

Ginsberg, J., Mohebbi, M.H., Patel, R.S., Brammer, L., Smolinski, M.S., Brilliant, L., 2009. Detecting influenza epidemics using search engine query data. Nature 457, 1012–U1014.

Goetz, T.B., Knetsch, T.A., 2019. Google data in bridge equation models for german GDP. Int. J. Forecast. 35, 45–66.

Guo, J.-F., Ji, Q., 2013. How does market concern derived from the Internet affect oil prices? Appl. Energy 112, 1536–1543.

Hajirahimi, Z., Khashei, M., 2019. Hybrid structures in time series modeling and forecasting: A review. Eng. Appl. Artif. Intell. 86, 83–106.

Han, L., Lv, Q., Yin, L., 2017. Can investor attention predict oil prices? Energy Econ. 66, 547–558.

Hou, A., Suardi, S., 2012. A nonparametric GARCH model of crude oil price return volatility. Energy Econ. 34, 618–626.

Huang, L., Wang, J., 2018. Global crude oil price prediction and synchronization based accuracy evaluation using random wavelet neural network. Energy 151, 875–888.

Huang, G.-B., Zhu, Q.-Y., Siew, C.-K., 2006. Extreme learning machine: Theory and applications. Neurocomputing 70, 489–501.

Kanungo, T., Mount, D.M., Netanyahu, N.S., Piatko, C.D., Silverman, R., Wu, A.Y., 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell. 24, 881–892.

Li, X., Ma, J., Wang, S., Zhang, X., 2015a. How does google search affect trader positions and crude oil prices? Econ. Model. 49, 162–171.

Li, X., Shang, W., Wang, S., 2019. Text-based crude oil price forecasting: A deep learning approach. Int. J. Forecast. 35, 1548–1560.

Li, X., Shang, W., Wang, S., Ma, J., 2015b. A MIDAS modelling framework for chinese inflation index forecast incorporating google search data. Electron. Comm. Res. Appl. 14, 112–125.

Liu, Z., Loo, C.K., Pasupa, K., Seera, M., 2020. Meta-cognitive recurrent kernel online sequential extreme learning machine with kernel adaptive filter for concept drift handling. Eng. Appl. Artif. Intell. 88.

Lu, Q., Li, Y., Chai, J., Wang, S., 2020. Crude oil price analysis and forecasting: A perspective of new triangle. Energy Econ. 87, http://dx.doi.org/10.1016/j.eneco.2020.104721.

Miao, H., Ramchander, S., Wang, T., Yang, D., 2017. Influential factors in crude oil price forecasting. Energy Econ. 68, 77–88.

Mohammadi, H., Su, L., 2010. International evidence on crude oil price dynamics: Applications of ARIMA-GARCH models. Energy Econ. 32, 1001–1008.

Moshiri, S., Foroutan, F., 2006. Forecasting nonlinear crude oil futures prices. Energy J. 27, 81–95.

Movagharnejad, K., Mehdizadeh, B., Banihashemi, M., Kordkheili, M.S., 2011. Forecasting the differences between various commercial oil prices in the Persian Gulf region by neural network. Energy 36, 3979–3984.

Ou, B., Zhang, X., Wang, S., 2012. How does China's macro-economy response to the world crude oil price shock: A structural dynamic factor model approach. Comput. Ind. Eng. 63, 634–640.

Panigrahi, S., Behera, H.S., 2017. A hybrid ETS ann model for time series forecasting. Eng. Appl. Artif. Intell. 66, 49–59.

Safari, A., Davallou, M., 2018. Oil price forecasting using a hybrid model. Energy 148, 49–58.

Scholkopf, B., Smola, A., Muller, K.R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 10, 1299–1319.

Shin, H., Hou, T., Park, K., Park, C.-K., Choi, S., 2013. Prediction of movement direction in crude oil prices based on semi-supervised learning. Decis. Support Syst. 55, 348–358.

Singleton, K.J., 2014. Investor flows and the 2008 boom/bust in oil prices. Manag. Sci. 60, 300–318.

Smith, P., 2016. Google's MIDAS touch: Predicting UK unemployment with internet search data. J. Forecast. 35, 263–284.

Song, T.M., Song, J., An, J.-Y., Hayman, L.L., Woo, J.-M., 2014. Psychological and social factors affecting internet searches on suicide in Korea: A big data analysis of google search trends. Yonsei Med. J. 55, 254–263.

Sun, S., Qiao, H., Wei, Y., Wang, S., 2017. A new dynamic integrated approach for wind speed forecasting. Appl. Energy 197, 151–162.

Sun, W., Sun, J., 2017. Daily PM2.5 concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm. J. Environ. Manag. 188, 144–152.

Sun, S., Wang, S., Wei, Y., Zhang, G., 2020. A clustering-based nonlinear ensemble approach for exchange rates forecasting. IEEE Trans. Syst. Man Cybern. S 50, 2284–2292.

Sun, S., Wei, Y., Tsui, K.-L., Wang, S., 2019. Forecasting tourist arrivals with machine learning and internet search index. Tourism Manag. 70, 1–10.

Tang, L., Wu, Y., Yu, L., 2018. A non-iterative decomposition-ensemble learning paradigm using RVFL network for crude oil price forecasting. Appl. Soft Comput. 70, 1097–1108.

Wang, J., Athanasopoulos, G., Hyndman, R.J., Wang, S., 2018a. Crude oil price forecasting based on internet concern using an extreme learning machine. Int. J. Forecast. 34, 665–677.

Wang, J., Li, X., Hong, T., Wang, S., 2018b. A semi-heterogeneous approach to combining crude oil price forecasts. Inform. Sci. 460, 279–292.

Wang, S., Yu, L., Lai, K.K., 2005. Crude oil price forecasting with TEI@I methodology. J. Syst. Sci. Complex 18, 145–166.

Wei, Y., Wang, Y., Huang, D., 2010. Forecasting crude oil market volatility: Further evidence using GARCH-class models. Energy Econ. 32, 1477–1484.

Yan, J., Zhang, B.Y., Liu, N., Yan, S.C., Cheng, Q.S., Fan, W.G., Yang, Q., Xi, W.S., Chen, Z., 2006. Effective and efficient dimensionality reduction for large-scale and streaming data preprocessing. IEEE Trans. Knowl. Data Eng. 18, 320–333.

Yu, L., Dai, W., Tang, L., 2016. A novel decomposition ensemble model with extended extreme learning machine for crude oil price forecasting. Eng. Appl. Artif. Intell. 47, 110–121.

Yu, L., Lai, K.K., Wang, S., 2008a. Multistage RBF neural network ensemble learning for exchange rates forecasting. Neurocomputing 71, 3295–3302.

Yu, L., Wang, S., Lai, K.K., 2008b. Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. Energy Econ. 30, 2623–2635.

Yu, L., Wang, Z., Tang, L., 2015. A decomposition-ensemble model with data-characteristic-driven reconstruction for crude oil price forecasting. Appl. Energy 156, 251–267.

Yu, L., Xu, H., Tang, L., 2017. LSSVR Ensemble learning with uncertain parameters for crude oil price forecasting. Appl. Soft Comput. 56, 692–701.

Zhang, Y., Ma, F., Wang, Y., 2019a. Forecasting crude oil prices with a large set of predictors: Can LASSO select powerful predictors? J. Empir. Financ. 54, 97–117.

Zhang, Y., Wei, Y., Zhang, Y., Jin, D., 2019b. Forecasting oil price volatility: Forecast combination versus shrinkage method. Energy Econ. 80, 423–433.

Zhong, X., Enke, D., 2017. Forecasting daily stock market return using dimensionality reduction. Expert. Syst. Appl. 67, 126–139.